

Leveraging Existing Bibliographic Metadata to Improve Automatic Document Identification in Web Archives

Mark Phillips, Cornelia Caragea, Praneeth Rika
IIPC Web Archiving Conference
May 11, 2023

Overview

Background

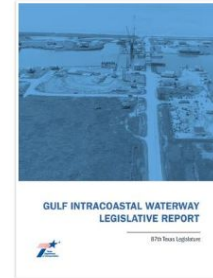
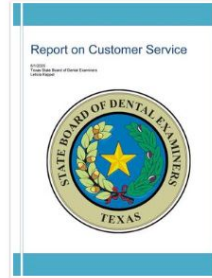
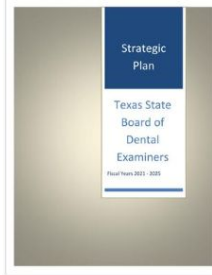
2017 Grant Overview

2022 Grant Overview

Work to date

Next Steps

Latest Additions



VIEW ALL

NDIIPP - Web At Risk

Research project in 2004

Interviews with web archivists about their collecting processes and workflows.

One line of discussion stuck, “We often collect web sites so that we can make sure and get the documents that are being placed on the web”

Archiving the web with the goal of collecting government publications so that they won't be lost and can be worked with in the future.



Collection Planning Guidelines

May 31, 2006

Prepared by:

Kathleen R. Murray
University of North Texas
krmurray@unt.edu

Inga K. Hsieh
University of North Texas
ikh0003@unt.edu

End of Term Experiments

In 2008 the End of Term Web Archive began to archive the US Federal .gov and .mil websites once every four years as part of our presidential election cycle.

In 2012 we started to analyze the PDF content in these collections and as you can imagine, identified millions of PDF files. In 2008 we identified 4.5 million unique PDF files.

Many of these would be wonderful additions to our Government Documents Collections.

But there were so many to sort through

Improving Access to Web Archives through Innovative Analysis of PDF Content

Mark Phillips and Kathleen Murray; University of North Texas Libraries; Denton, Texas, USA

Abstract

In 2008 five United States institutions collaborated to archive the U.S. federal government Web presence: the Library of Congress, the Internet Archive, the California Digital Library, the Government Printing Office, and the University of North Texas (UNT). Their objective was to document the changes coincident with the shift in leadership of the U.S. executive branch. The five partners identified key resources from the U.S. .gov Top Level Domain and completed crawls from September 2008 until March 2009. The resulting End of Term (EOT) 2008 Web Archive, a 16 TB dataset, was distributed to partners interested in providing local services and access to the archive. The UNT Libraries investigated Portable Document Format (PDF) files, a class of content many information professionals associate with the traditional notion of "discrete documents". Over four million unique PDF documents were extracted from the Archive and a series of metadata and information extraction processes were conducted for each document. Additionally, derivative raster images of the first page of each document were created. These metrics were ingested into a database for further analysis, which brought to light previously hidden characteristics of the federal government's Web-published content. The paper discusses the overall workflow and describes the tools used to extract document features. Findings suggest opportunities for the development of retrieval tools that will provide new ways of selecting content and building collections from large Web archives.

Background

As Web archives become more available, organizations will seek to include materials from these repositories in their collections. However, such inclusion is often precluded by content identification and selection challenges. This is in part because the high-level metadata associated with Web archive files does not support material selection in a manner consistent with libraries' collection development policies. To address this problem, the University of North Texas (UNT) Libraries conducted a needs assessment in 2005 as a part of the Web-at-Risk project, a digital preservation project of the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP) [1]. The study identified collection development needs and issues confronting librarians, archivists, content providers, and researchers who deal with the challenges posed by changes in the publication and distribution of U.S. government information. A number of government information professionals identified the PDF format as being of significance in their collection development processes. In fact, for many professionals PDF-formatted documents were the unit they were most interested in capturing during the Web archiving process [2].

In 2009, UNT Libraries received a research grant from the Institute of Museum and Library Services (IMLS) to continue investigating libraries' collection development needs relative to Web-published government information (Classification of the End of Term Archive Project; IMLS LG-06-09-0174-09) [3]. UNT leveraged its participation in the End of Term Web Archive (EOT 2008 Archive) project, a collaborative effort of the Library of Congress, the Internet Archive, the California Digital Library, the U.S. Government Printing Office, and the University of North Texas [4]. This important project captured the entirety of the federal government's public Web presence before and after the 2009 change in U.S. presidential administrations. The result is the 16-terabyte EOT 2008 Archive containing 160,211,356 URLs [5]. The largest Top Level Domains (TLDs) are listed in Table 1 and the top four file formats by number of mime-type are listed in Table 2.

Table 1. Number of URLs & Subdomains by Top Level Domains

Top Level Domains	# URLs	# Unique Sub-domains
.gov	137,780,023	14,338
.com	7,805,205	57,873
.org	5,107,552	29,798
.mil	3,554,956	1,877
.edu	3,551,845	13,856

The UNT Libraries was interested in providing government information professionals with mechanisms to identify resources of interest for their collections within the very large, and relatively inaccessible, EOT 2008 Archive. Because of the previously documented interest of government information professionals in archived PDF documents, as well as the fact that over 10 million PDF documents are represented in the Archive, the PDF files were a logical subset of content to investigate in a systematic manner. The project team sought to improve its understanding of this important class of content.

The overarching question directing this investigation was: Is it feasible to describe the content of Web archives by format-specific features? If so, it may also be feasible to take advantage of the descriptive findings and use them to inform the development of mechanisms that aid information professionals in their collection building processes.

End of Term Publications

During the 2016 End of Term project we identified all of the PDF documents that had been nominated for capture.

These totalled over 1,900.

We extracted these from our crawls and built a digital collection for these in the UNT Digital Library

We worked with volunteers to create metadata records these documents so they could be easily accessed.

University Libraries
UNT Digital Library

HOME COLLECTIONS PARTNERS TITLES LOCATIONS TYPES DATES ABOUT TOUR CONTACT US

University Libraries / UNT Digital Library / Explore / Collections / End of Term Publications

End of Term Publications

The End of Term Publications collection consists of reports, presentations, and documents collected during one of the End of Term Presidential Web Archive projects either in 2008, 2012, or 2016. Items were either explicitly nominated for inclusion in the EOT archive or have been extracted from the EOT Archive for inclusion in this collection.

For more information about the End of Term Web Archive see the homepage at <http://eotarchive.cdlib.org>

Search inside this Collection Other Search Options ▾

Look In: Full Text

At a Glance

1,945 Items	17 Types	135 Titles
2 Partners	5 Decades	2 Languages
24 Counties	50 States	41 Countries
332,456 Usage	6 years, 3 months ago Collection Created	3 years, 9 months ago Last Updated

<https://digital.library.unt.edu/explore/collections/EOT/>

CyberCemetery Extracted Publications

Many of the websites archived in the CyberCemetery existed as a way of publishing a final report that was also submitted to Congress.

These reports are present in the web archive but users had to know how to look for them.

A clear improvement to the user experience is to make these publications standard items in the digital library with proper metadata for discovery.

It might seem like an obvious thing to do but didn't cross our mind for an embarrassingly long time.

The screenshot shows the UNT Digital Library website. The header includes the UNT logo and navigation links: HOME, COLLECTIONS, PARTNERS, TITLES, LOCATIONS, TYPES, DATES, ABOUT, TOUR, CONTACT US. The main content area is titled "CyberCemetery Extracted Publications" and features a description: "This growing collection of reports and other publications comes from defunct government websites preserved in our CyberCemetery. The items were automatically identified using machine learning algorithms developed under the auspices of an Institute of Museum and Library Services (IMLS) National Digital Platform research grant." Below the description is a search bar with the text "Search inside this Collection" and a "Search" button. The "At a Glance" section displays a grid of statistics:

162 Items	5 Types	3 Titles
1 Partner	4 Decades	1 Language
8 Countries	9 States	2 Countries
28,971 Usage	4 years, 7 months ago Collection Created	1 year, 6 months ago Last Updated

<https://digital.library.unt.edu/explore/collections/GDCCP/>

2017 IMLS Grant Project

Explore the use of machine learning models to identify and classify “in scope” publications that exist in web archives.

Three domains of experiment

- State Publications
- Federal Technical Reports
- University Faculty Publications

Overall work was successful with models able to correctly identify in scope publications.

Challenges often resulted in not enough labelled data for more advanced models.

The screenshot shows the IMLS website interface. At the top, there is a navigation menu with links for About, Grants, Our Work, Data, News, and Contact. The main content area displays the grant details for LG-71-17-0202-17, awarded to the University of North Texas. The grant information includes the program (National Leadership Grants - Libraries), fiscal year (2017), federal funds amount (\$318,988), city (Denton), and state (TX). A description of the project is provided, detailing the research on machine-learning algorithms for identifying content-rich PDF and Word documents in web archives. Below the description, there is a section for Project Proposals with a table of attachments.

Attachment	Size
Ig-71-17-0202-17-full-proposal-documents.pdf	397.19 KB
Ig-71-17-0202-17-preliminary-proposal.pdf	107.09 KB

<https://www.ims.gov/grants/awarded/Ig-71-17-0202-17>

2017 Grant Cont.

In addition to experiments with different approaches, we created datasets that could be used by others to experiment.

We also conducted qualitative research with a dozen web archiving and collection professionals to understand how they select.

A major finding was that existing library catalogs are often referenced as containing the “collecting history” of an organization.

This was especially true for state government documents collections.

The screenshot displays the UNT Digital Library interface. The header includes the UNT logo and navigation links: HOME, COLLECTIONS, PARTNERS, TITLES, LOCATIONS, TYPES, DATES, ABOUT, TOUR, CONTACT US. The main content area features a search bar with the query 'fox phillips tarver' and 1 matching result. The document title is 'Programmatic Extraction of Documents from Web Archives: Identifying Document Characteristics from Content Selector Interviews'. The description states it is a 25-page paper documenting interviews with professionals managing collections. The authors are Fox, Nathaniel T.; Phillips, Mark Edward & Tarver, Hannah 2020. The page includes a 'Mapped' button, a search icon, and an 'Open Access' button. A PDF version is also available for download.

<https://digital.library.unt.edu/ark:/67531/metadc1757659/>

IMLS 2022 Grant Project

Leveraging Existing Bibliographic Metadata to Improve Automatic Document Identification in Web Archives

Extension of the research from 2017, specifically can we leverage bibliographic metadata from library catalogs and digital collections to build better models for document classification.

Can we use these models to reduce the human labor involved in building larger labelled datasets for training.

What metadata is most useful for this kind of model building.

The screenshot shows the IMLS website interface. At the top, there is a navigation menu with links for About, Grants, Our Work, Data, News, and Contact. The main content area displays the following information:

- Program:** National Leadership Grants - Libraries
- Fiscal Year:** 2022
- Federal Funds:** \$385,769
- City:** Denton
- State:** TX

The grant title is "University of North Texas (University of North Texas, University Libraries)" with Log Number "LG-252349-OLS-22". A description states: "University of North Texas Libraries, partnering with the University of Illinois Chicago (UIC) Computer Science Department, will conduct a research project with the long-term objective of improving access to digital resources housed in web archives. The project team will investigate the potential of using existing bibliographic metadata related to state government document collections to better train machine learning models that can assist librarians and information professionals in identifying and classifying high-value publications from large web archives. The project team will share all datasets, algorithms, and tools resulting from this project through GitHub and the project webpage, and they will communicate research findings through publications and presentations at conferences on library science, information retrieval, artificial intelligence, and natural language processing. Subrecipient, UIC, will be responsible for the machine learning component of the project and will help disseminate research findings."

Under the heading "Project Proposals", there is a table with two columns: "Attachment" and "Size".

Attachment	Size
LG-252349-OLS-22 Full Proposal	541.41 KB
LG-252349-OLS-22 Preliminary Proposal	195.66 KB

At the bottom of the page, there is a "CONTACT US" section with the phone number 202-653-4657 and email imlsinfo@imls.gov. The footer contains various links: Viewers & Players, FOIA, No FEAR, Privacy & Terms of Use, EEO, Accessibility, Open Government, Office of Special Counsel, and USA.gov.

<https://www.imls.gov/grants/awarded/lg-252349-ols-22>

Research Questions

Project Goal: The overarching goal of this project is to investigate the potential of using existing bibliographic metadata related to state government document collections to better train machine learning models that can assist librarians and information professionals in identifying and classifying high-value publications from large web archives.

Research Questions:

1. How can large amounts of training data be generated for supervised approaches with less intensive human effort, which is often impractical?
2. How can we successfully incorporate information from unlabeled data to build robust classifiers for identifying documents in-scope of a collection?
3. How will our models generalize to data “in the wild” (i.e., data from a different state) and how robust are the models under distribution or vocabulary shifts (e.g., on data from one state to another under vocabulary distribution shifts, or from one collection type/scope to another), when no human-annotated datasets are available in the new / target domain?

Grant Overview

Collaboration between the UNT Libraries and the Department of Computer Science at the University of Illinois Chicago

External data collaborators are the Library of Michigan and Archive It.

Advisory board of experts in metadata, web archiving, state publications, government information and machine learning.

Project team includes two primary investigators and two graduate research assistants.

The screenshot displays the 'The Portal to Texas History' website interface. The main content area is titled 'Texas State Publications' and includes a description: 'This growing collection of materials produced by the State of Texas includes agency annual reports, legislative publications, statistical reports, and various state government reports and periodicals.' Below this is a search bar with the text 'Search Inside this Collection' and a 'Search' button. A navigation menu on the left includes 'About this Collection', 'Overview', 'At a Glance', 'Latest Additions', 'Cite This Collection', 'Explore Holdings', and 'Contact Us'. A 'Share' section with social media icons is also present. The 'At a Glance' section features a grid of statistics:

19,505 Items	21 Types	1373 Titles
4 Partners	15 Decades	6 Languages
268 Countries	29 States	13 Countries
3,176,879 Usage	10 years, 3 months ago Collection Created	3 days, 6 hours ago Last Updated

<https://texashistory.unt.edu/explore/collections/TXPUB/>

Grant Data Scope

Texas State Publications

- 19,500+ records from a digital collection in The Portal to Texas History
- 13,785 records from our library catalog related to Texas Government Documents
- texas.gov web archive from 2012 and also 2023.

Michigan State Publications

- 5,439 records from their Digital Publications Collection
- MARC records from their library system
- michigan.gov collection from Archive-It

The screenshot displays the UNT Discover library catalog interface. The header includes the UNT logo and navigation links: HOME, ABOUT, RESEARCH, SERVICES, SPACES, NEWS, CALENDAR, ASK US, ACCOUNTS, LOGIN. The search bar shows 'Your Search Terms:' and 'Look in: All Fields'. The search results are filtered by 'Collection > Government Documents' and 'Subject - Region > Texas'. The results list includes:

- 9-1-1 caller [1980s to present]**: Texas Advisory Commission on State Emergency Communications. Available - Gov Docs Storage. More available.
- The Advisor: an official publication of the Texas Real Estate Commission [1950 to present]**: Texas Real Estate Commission. Available - Gov Docs Storage. More available.
- Aranas, National Wildlife Refuge visitor information and map [20th century to present]**: U.S. Fish and Wildlife Service. Available - Syncmore.

<https://discover.library.unt.edu/>

Grant Activity to Date

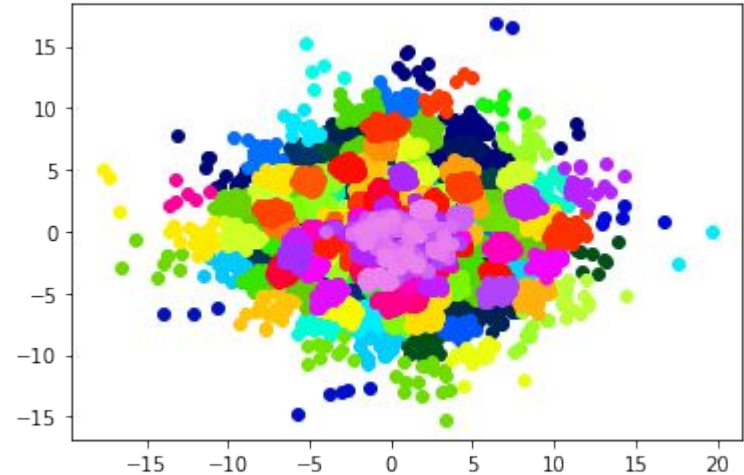
Exploring state web archives and government publications repositories

- 48/50 have recognizable digital publications collections
- 34/50 have some web archiving activity

Working on building datasets

Exploring metadata from state publications repositories from other states

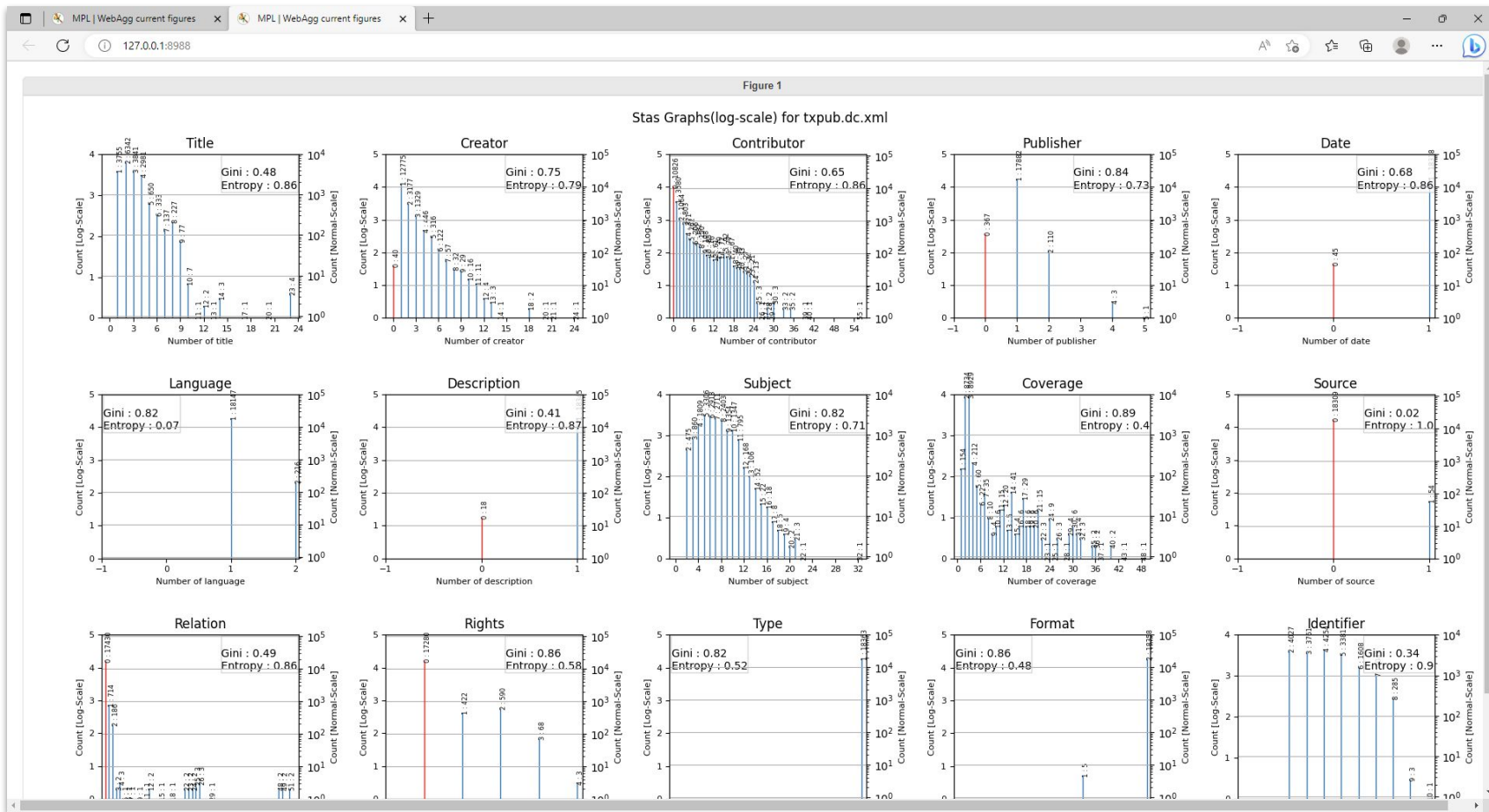
Visualizing and describing collections of metadata from these repositories.



Texas State Publications - The Portal to Texas History

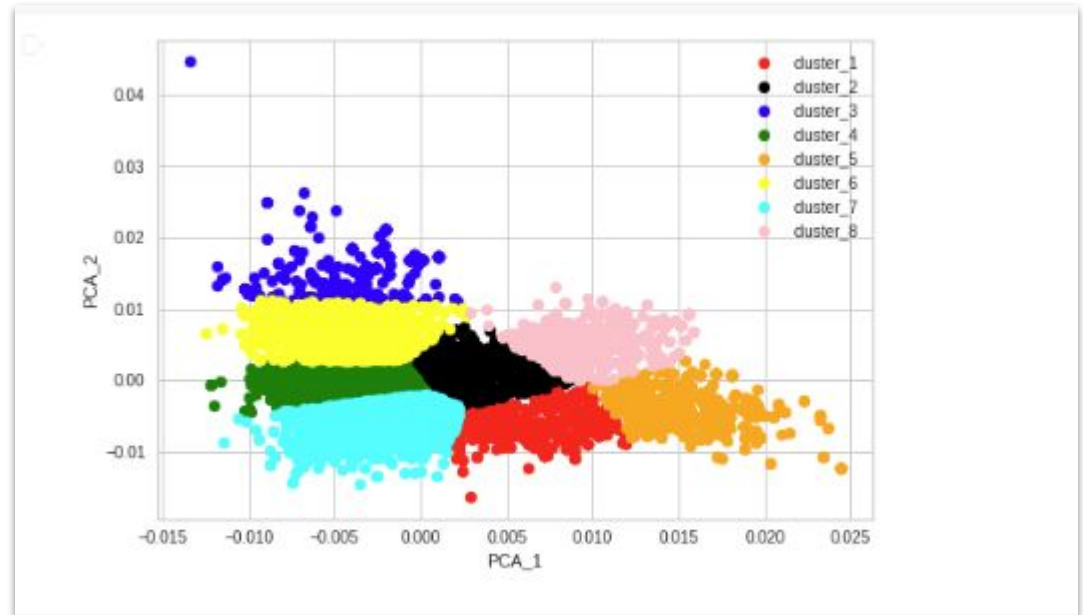
Element Name	Records with Element Instances	Percentage of Records with Element Instances	Unique data values in Element Instances	Mean Instances per record	Mode Instances per record	Frequency of Mode Instances per record	Entropy
title	18363	100.0%	25221	2	2	34.54%	0.864
creator	18323	99.78%	4923	1	1	69.57%	0.794
contributor	7537	41.04%	7909	3	1	19.5%	0.861
publisher	17996	98.0%	928	1	1	97.38%	0.727
date	18318	99.75%	4343	1	1	99.75%	0.859
language	18363	100.0%	6	1	1	98.82%	0.069
description	18345	99.9%	10682	1	1	99.9%	0.870
subject	18363	100.0%	17539	6	5	18.0%	0.710
coverage	18363	100.0%	3905	2	3	48.62%	0.400
source	54	0.29%	53	1	1	0.29%	0.998
relation	933	5.08%	855	1	1	3.89%	0.863
rights	1083	5.9%	110	1	2	3.21%	0.579
type	18363	100.0%	21	1	1	100.0%	0.516
format	18363	100.0%	4210	1	2	99.97%	0.483
identifier	18363	100.0%	47785	3	4	23.17%	0.902
Table : 1 Texas State Collection Basic Stats							

Texas State Publications - The Portal to Texas History



Texas State Publications - The Portal to Texas History

Clustering metadata records using vectors created with word embeddings

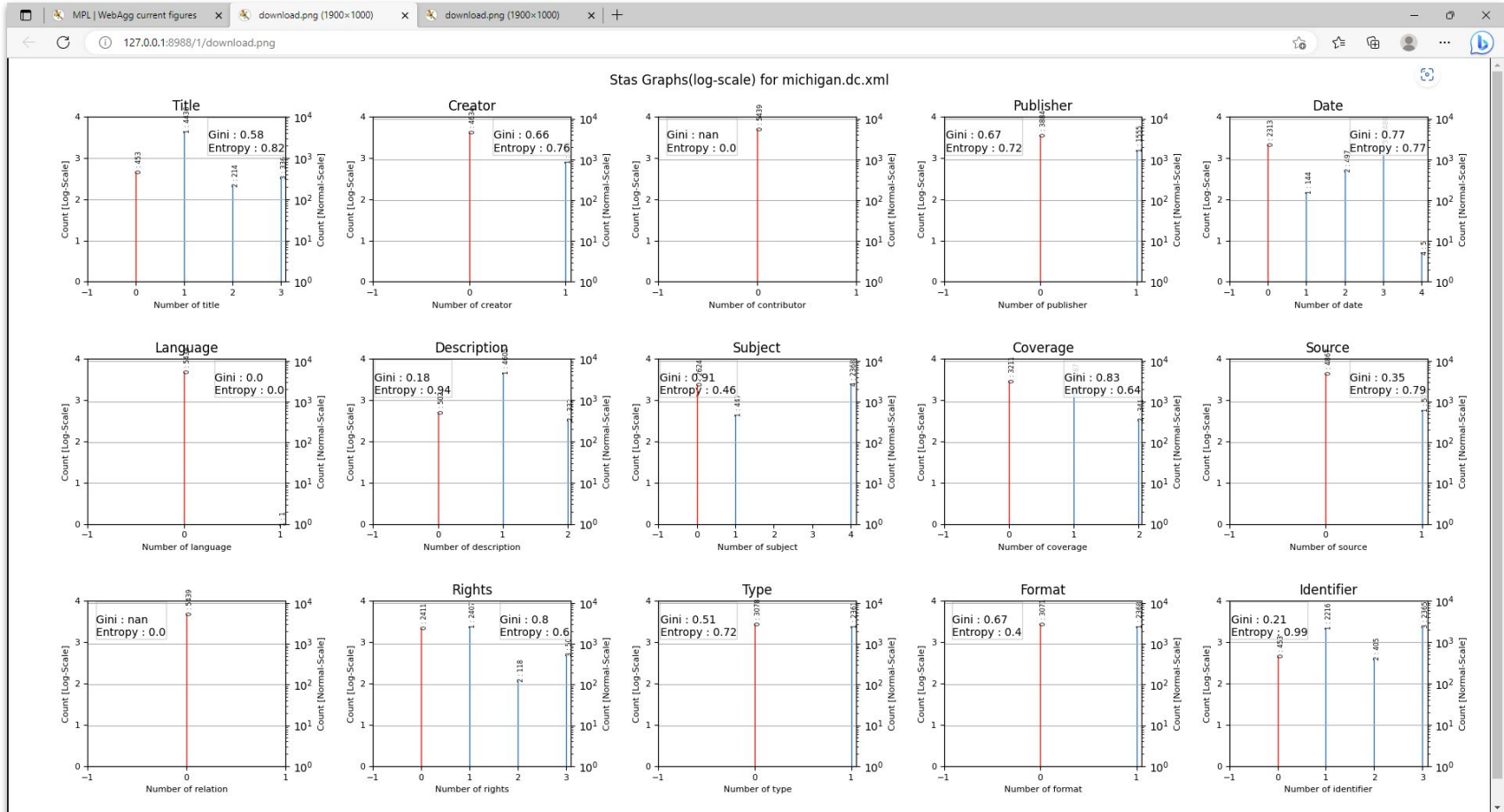


Library of Michigan Digital Collection

Element Name	Records with Element Instances	Percentage of Records with Element Instances	Unique data values in Element Instances	Mean Instances per record	Mode Instances per record	Frequency of Mode Instances per record	Entropy
title	4986	100.0%	2259	1	1	88.97%	0.823
creator	805	16.15%	201	1	1	16.15%	0.753
contributor	0	0%	0	0	0	0%	0
publishe	1555	31.19%	406	1	1	31.19%	0.722
date	3126	62.7%	1319	2	3	49.74%	0.773
language	1	0.02%	1	1	1	0.02%	0
description	4937	99.02%	4281	1	1	92.36%	0.937
subject	2815	56.46%	369	3	4	47.49%	0.463
coverage	2228	44.69%	205	1	1	37.85%	0.640
source	572	11.47%	2	1	1	11.47%	0.404
relation	0	0%	0	0	0	0%	0
rights	3028	60.73%	31	1	1	48.28%	0.598
type	2361	47.35%	7	1	1	47.35%	0.722
format	2368	47.49%	5	1	1	47.49%	0.401
identifier	4986	100.0%	7414	2	3	47.43%	0.988
Table : 1 Michigan State Collection Basic Stats							

Total_records : 5439
 Deleted_records : 453
 Available_records : 4986

Library of Michigan Digital Collection



Recreating previous grant work with current toolkits and workflows.

Moving from TensorFlow to PyTorch

Creating baseline implementations for future experiments



Labeled PDF Dataset from Texas Records and Information Locator (TRAIL) Web Archive

This dataset contains a random sample of 2000 PDF documents from the Texas Records and Information Locator (TRAIL) Web Archive from the Texas State Library and Archives Commission. Each PDF has been sorted into two categories, TX_Pub_In_Scope and Not_TX_Pub.

DATE: July 2018

CREATOR: Tarver, Hannah & Phillips, Mark Edward

PARTNER: UNT Libraries



The Portal to Texas History's Texas State Publications Collection Dataset

This dataset contains a set of 2,448 PDF files from the Texas State Publications collection in The Portal to Texas History.

DATE: September 12, 2018

CREATOR: Phillips, Mark Edward

PARTNER: UNT Libraries

[199] pdf

	file_name	title	text	layout	page_count	target	file_type	file_size	total_words	word_count	selected_words
0	24LOI5BAD2PKTQ6SUMPFNPTATFJCM	Vaccine Billboard Ad	It takes more than a kiss.\n\nP+ walelelTinal...	792.0 * 360.0	1	TX_Pub_In_Scope	text_pamphlet	1486421	[[takes, kiss, walelelTinal, build, child s,...	10	[takes, kiss, walelelTinal, build, child s, ...
1	252QHHKHPARPIPSWWHPFC5MTWPS5TF42	Microsoft PowerPoint - ACS Chartbook 2006_FINAL	Acknowledgements\nThis report was researched a...	612.0 * 792.0	34	TX_Pub_In_Scope	text_report	1070197	[[acknowledgements, report, researched, writte...	2128	[acknowledgements, report, researched, written...
2	276MTABTD4CBWDCFO2Y4USPLGYZLJOPF	NaN	Joint Semi-Annual Interagency Coordination Rep...	612.0 * 792.0	10	TX_Pub_In_Scope	text_report	56801	[[joint, semi annual, interagency, coordinatio...	1203	[joint, semi annual, interagency, coordinatio...
3	2HW4GGGX2DVM6P56L32L43LPSKMBJU	Implementation of Arlington Ramp Metering System	Project Summary Report 3982-S\nProject 7-3982:...	621.7200317382812 * 776.6640014648438	4	TX_Pub_In_Scope	text_report	1739791	[[project, summary, report, fort, worth, real ...	479	[project, summary, report, fort, worth, real t...
4	2HWP7LUEJR7EQ774BSZUDT3Z6BNEXP4M	Legislative Report	B78-1231-1M-L\n\nJOURNAL\n\nof THE\n\nSENATE O...	595.0 * 842.0	219	TX_Pub_In_Scope	text_leg	42470241	[[journal, senate, texas, second, called, sess...	6705	[journal, senate, texas, second, called, sessi...
...
130	YVS63FMDFTRYU427QIHG3JYRE3KWC56G	Rusk County Groundwater	\nRusk County Groundwater\n\nConservation Dist...	612.0 * 792.0	24	TX_Pub_In_Scope	text_pamphlet	1073297	[[rusk, county, groundwater, conservation, dis...	1797	[rusk, county, groundwater, conservation, dist...

0s completed at 3:06 PM

Grant Next Steps

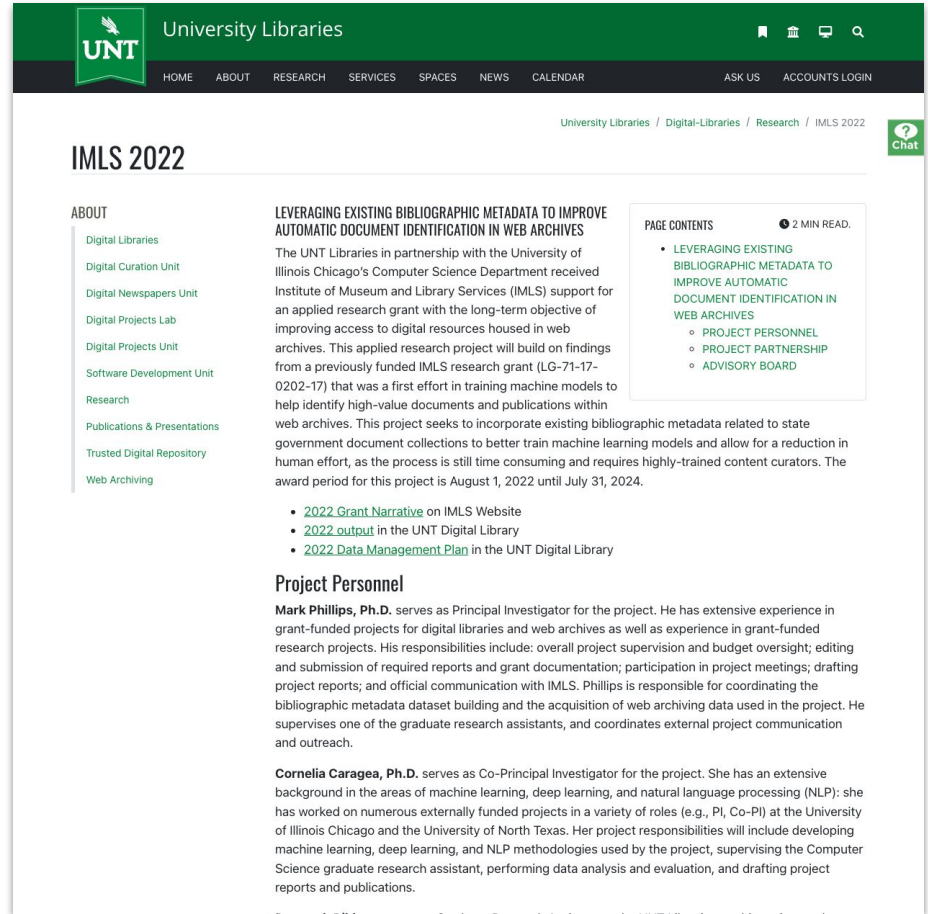
UNT is working on packaging Texas and Michigan Datasets for distribution.

Machine learning track of project will begin this summer at UIC

Begin to reimplement and share workflows and tools for replicating work.

<https://library.unt.edu/digital-libraries/research/ims-2022/>

<https://github.com/state-pubs-from-web-archives>



University Libraries

HOME ABOUT RESEARCH SERVICES SPACES NEWS CALENDAR ASK US ACCOUNTS LOGIN

University Libraries / Digital-Libraries / Research / IMLS 2022

IMLS 2022

ABOUT

- Digital Libraries
- Digital Curation Unit
- Digital Newspapers Unit
- Digital Projects Lab
- Digital Projects Unit
- Software Development Unit
- Research
- Publications & Presentations
- Trusted Digital Repository
- Web Archiving

LEVERAGING EXISTING BIBLIOGRAPHIC METADATA TO IMPROVE AUTOMATIC DOCUMENT IDENTIFICATION IN WEB ARCHIVES

The UNT Libraries in partnership with the University of Illinois Chicago's Computer Science Department received Institute of Museum and Library Services (IMLS) support for an applied research grant with the long-term objective of improving access to digital resources housed in web archives. This applied research project will build on findings from a previously funded IMLS research grant (LG-71-17-0202-17) that was a first effort in training machine models to help identify high-value documents and publications within web archives. This project seeks to incorporate existing bibliographic metadata related to state government document collections to better train machine learning models and allow for a reduction in human effort, as the process is still time consuming and requires highly-trained content curators. The award period for this project is August 1, 2022 until July 31, 2024.

- [2022 Grant Narrative](#) on IMLS Website
- [2022 output](#) in the UNT Digital Library
- [2022 Data Management Plan](#) in the UNT Digital Library

Project Personnel

Mark Phillips, Ph.D. serves as Principal Investigator for the project. He has extensive experience in grant-funded projects for digital libraries and web archives as well as experience in grant-funded research projects. His responsibilities include: overall project supervision and budget oversight; editing and submission of required reports and grant documentation; participation in project meetings; drafting project reports; and official communication with IMLS. Phillips is responsible for coordinating the bibliographic metadata dataset building and the acquisition of web archiving data used in the project. He supervises one of the graduate research assistants, and coordinates external project communication and outreach.

Cornelia Caragea, Ph.D. serves as Co-Principal Investigator for the project. She has an extensive background in the areas of machine learning, deep learning, and natural language processing (NLP): she has worked on numerous externally funded projects in a variety of roles (e.g., PI, Co-PI) at the University of Illinois Chicago and the University of North Texas. Her project responsibilities will include developing machine learning, deep learning, and NLP methodologies used by the project, supervising the Computer Science graduate research assistant, performing data analysis and evaluation, and drafting project reports and publications.

PAGE CONTENTS 2 MIN READ.

- LEVERAGING EXISTING BIBLIOGRAPHIC METADATA TO IMPROVE AUTOMATIC DOCUMENT IDENTIFICATION IN WEB ARCHIVES
 - PROJECT PERSONNEL
 - PROJECT PARTNERSHIP
 - ADVISORY BOARD

Thank you.

mark.phillips@unt.edu

@vphill